JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Le preferenze di selezione verbali: un approccio computazionale

Raffaele Guarasci

Introduzione

L'esperimento effettuato e presentato in questo lavoro mira a fornire una rappresentazione delle preferenze di selezione verbali per la lingua italiana, utilizzando delle metodologie basate su corpora. Per preferenze di selezione si intendono i vincoli imposti dai predicati nella realizzazione dei propri argomenti. L'acquisizione di preferenze di selezione da un corpus, proposta inizialmente da Philip Resnik nel 1993 («Semantic classes and syntactic ambiguity»), si può articolare in due fasi: l'estrazione degli argomenti dai corpora scelti e la generalizzazione delle preferenze di selezione verbali da un'ontologia lessicale.

In questo lavoro vengono utilizzati:

LexIt, lessico di valenza per i verbi della lingua italiana come risorsa lessicale;

MultiWordNet, database lessicale multilingua strutturato in classi semantiche organizzate in modo gerarchico, come ontologia (Magnini e Strapparava).

JLIS.it. Vol. 3, n. 1 (Giugno/June 2012).
DOI: 10.4403/jlis.it-4786

L'analisi proposta si ricollega a uno degli ultimi modelli sviluppati per generalizzare le preferenze di selezione, l'esperimento effettuato da Sabine Schulte Im Walde («Experiments on the Automatic Induction of German Semantic Verb Classes») sui verbi tedeschi, mirato a ottenere una rappresentazione ad alto livello che generalizzi le preferenze di selezione degli argomenti e fornisca una distribuzione del comportamento dei verbi nella lingua tedesca. Il lavoro effettuato sulla lingua tedesca aveva come obiettivo quello di fornire una generalizzazione del comportamento dei verbi, considerando soltanto le classi più generiche e astratte in cima alla rete semantica. Questo alto livello di generalizzazione limita l'osservazione di alcuni comportamenti specifici di determinati verbi. L'analisi qui proposta mira quindi a superare tale problema e a fornire un livello di rappresentazione più dettagliato, che caratterizzi meglio specificità e preferenze di selezione, considerando nella navigazione della rete semantica di MultiWordNet tutte le classi intermedie, dal momento che le classi generali risultano troppo ampie per caratterizzare in modo differente tutti i verbi. Esprimere le preferenze di selezione nei termini di tutte le classi intermedie della gerarchia permette di far venire alla luce comportamenti più specifici nella scelta degli argomenti, che non potrebbero essere considerati guardando solo le classi generiche.

Le preferenze di selezione

Si può definire "preferenza di selezione" la proprietà di un verbo di preferire o meno argomenti di un particolare tipo semantico. Si può dunque vedere la preferenza di selezione come una sorta di vincolo che opera una restrizione specificando quale siano gli argomenti adatti per un dato predicato. Il concetto di preferenze di selezione o di restrizioni semantiche ha una lunga storia ed è

stato ampiamente trattato sia nella Linguistica generativa (Katz e Fodor; Chomsky), che nella Linguistica computazionale. Uno dei primi approcci per caratterizzare le preferenze di selezione è stato quello di Katz e Fodor, basato sulla più ampia teoria semantica della decomposizione del significato delle parole in features lessicali caratterizzanti. Il classico esempio proposto per la lingua inglese è la parola *bachelor* che può indicare un uomo non sposato, scapolo (features: maschio e umano) o un esemplare di foca maschio privo di compagna (features: maschio e animale). Applicare questo modello ai predicati significa identificare per i predicati delle condizioni necessarie e sufficienti perché siano semanticamente accettabili per essere associate a quell'argomento. Tali condizioni sono rappresentate come funzioni booleane. Questo approccio presenta diversi limiti: dalla difficoltà di identificare univocamente delle condizioni necessarie e sufficienti sempre valide, all'impossibilità di formulare alcuni concetti tramite features binarie.

Modello di Resnik

Il modello elaborato da Philip Resnik nel 1993 è considerato il più quotato modello computazionale per le preferenze di selezione e il punto di riferimento per i lavori successivi. La strategia adottata da Resnik si compone di una rappresentazione tassonomica dei concetti e di una formalizzazione probabilistica delle preferenze di selezione definite nei termini di quella tassonomia, che vengono poi computate e analizzate sulla base di frequenze di co-occorrenza tra i predicati e i loro argomenti. La prima componente del modello deve essere una tassonomia concettuale, una rete semantica, nella quale le classi di concetti devono essere strutturate. La seconda componente del modello deve, invece, fornire una caratterizzazione delle preferenze di selezione nei termini di una relazione probabilistica tra predicati e

classi concettuali, basandosi sull'assunzione che un predicato tende ad associarsi prevalentemente con determinate classi di argomenti.

Implementazione computazionale

Nella sua realizzazione computazionale, il metodo di Resnik utilizza come rappresentazione tassonomica l'ontologia di Wordnet,[1] inserendo nel modello tutti i synset rappresentati in WordNet (mentre altri approcci selezionano solo alcuni topnode, in modo da ottenere una rappresentazione più astratta delle preferenze semantiche). L'algoritmo di Resnik funziona nel seguente modo:

1. si assegna il conteggio di co-occorrenza ai synset di WordNet che contengono un nome come testa lessicale; quando una parola ricorre in più di un synset, la sua frequenza viene divisa per il numero di synset. Nelle varie simulazioni del modello, Resnik utilizza principalmente il *Brown Corpus of American English* (Francis e Kucera), il corpus di riferimento della lingua inglese, e CHILDES (MacWhinney e Snow), un corpus comprendente una serie di interazioni dialogiche con bambini;

2. si estende il conteggio di co-occorrenza a tutti i nodi della gerarchia di WordNet.

Questo metodo è utilizzato per calcolare la prior distribution, p(classe), cioè la probabilità che una classe di WordNet prenda una particolare posizione sintattica, indipendentemente dal predicato, che viene stimata usando soltanto la frequenza della classe costituita a partire dagli argomenti di un dato slot, e la posterior distribution,

[1] WordNet è un database semantico contenente 90.000 parole inglesi tra verbi, nomi, aggettivi e avverbi, in cui le parole sono organizzate in classi semantiche. Le parole che condividono un topnode nell'ontologia formano un synset. Un synset comprende un insieme di concetti legati da relazioni di sinonimia, meronimia o altro.

p(classe | predicato), cioè la probabilità di una classe di WordNet nella stessa posizione ma tenendo conto del predicato, stimata tramite la frequenza della coppia predicato-nome. La comparazione tra prior distribution e posterior distribution serve a quantificare quanto una classe semantica si adatti a uno slot predicato-argomento; le classi semantiche che si adattano meglio a un particolare slot hanno le probabilità più alte di co-occorrere con questi. Le preferenze di selezione sono quindi costituite dalla relazione tra prior distribution e la posterior distribution, ovvero tra la probabilità del verificarsi di un argomento a prescindere dal predicato e la probabilità condizionata tra i due. La figura 1 nella pagina seguente, tratta da Resnik («Selection and information: a class-based approach to lexical relationships»), illustra questo approccio: dato un set di classi, C (nell'esempio *Legume*, *Animal*, *Trait*), un predicato, *to grow*, e una posizione sintattica, oggetto diretto, la prior distribution p(classe) è confrontata con la posterior distribution p(classe | predicato). Tralasciando il predicato, p(class), *Animal* tende a ricorrere più frequentemente come oggetto diretto rispetto a *Legume*; tuttavia se si effettuano i calcoli introducendo i predicati, *Legume* diventa molto più frequente di *Animal*.

Modello Schulte Im Walde

Il modello elaborato da Sabine Schulte Im Walde per l'assegnazione di tipi semantici agli argomenti dei verbi della lingua tedesca è basato sul medesimo approccio proposto da Resnik. Per ogni combinazione verbo-frame-slot, la frequenza dei filler nominali è estesa alla gerarchia dei 15 topnode esclusivi di GermaNet (Hamp e Feldweg) (*Creature, Thing, Property, Substance, Food, Means, Situation, State, Structure, Body, Time, Space, Attribute, Cognitive Object, Cognitive Process*). La frequenze delle parole collegate a più di un concetto viene divisa uniformemente tra questi. La differenza principale

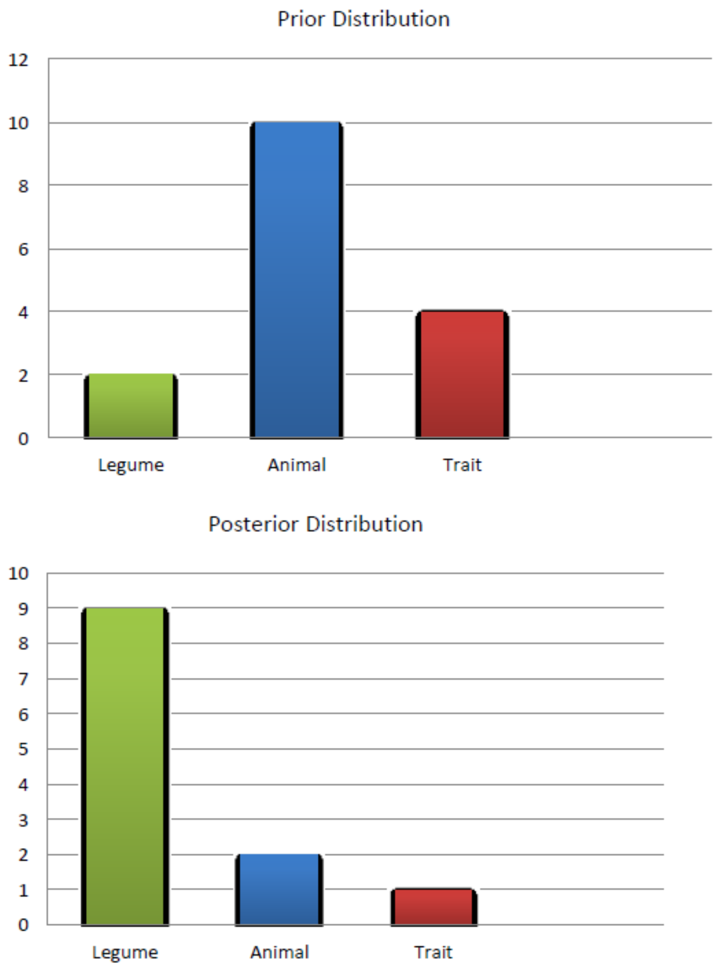

Prior Distribution
12
10
8
6
4
2
0
Legume
Animal
Trait
Posterior Distribution
10
9
8
7
6
5
4
3
2
1
0
Legume
Animal
Trait

Figura 1

rispetto all'algoritmo di Resnik consiste nel non utilizzare tutti i synset dell'ontologia di WordNet, ma soltanto questi 15 topnode esclusivi. Le frequenze risultanti dei 15 nodi che occorrono in ogni slot sono poi usate per definire una distribuzione probabilistica, normalizzata sulla frequenza totale di co-occorrenza verbo-frame. Oltre a considerare solo questa selezione di synset di GermaNet, l'agoritmo di Schulte Im Walde non apporta sostanziali differenze a Resnik nel calcolo della frequenza di co-occorrenza tra slot e classi, mentre per effettuare il clustering Schulte Im Walde utilizza il valore p(classe | predicato), diversamente da Resnik che usa una misura proveniente dalla teoria dell'informazione per caratterizzare in modo più ampio le associazioni tra classi semantiche e predicati.

LexIt

LexIt (Lenci, «Carving Verb Classes from Corpora») è una risorsa lessicale per lo studio e l'analisi dei verbi della lingua italiana costruita in modo completamente automatico e basata sul corpus *La Repubblica* e Wikipedia. LexIt rappresenta di fatto la prima risorsa per la lingua italiana che comprenda informazioni distribuzionali sul comportamento dei verbi italiani.[2] L'obiettivo che questo strumento si pone è quello di descrivere il comportamento dei verbi della lingua italiana sotto il profilo distribuzionale sintattico e semantico, estraendo le informazioni necessarie in modo automatico. il processo per raggiungere questo obiettivo si articola in diversi sottotask: estrazione dei frame di sottocategorizzazione, assegna-

[2]La risorsa lessicale più simile a LexIt è rappresentata dal lessico SIMPLE, il quale contiene informazioni sulle preferenze di selezione verbali per le classi semantiche, ma non per filler lessicali o polisemia degli argomenti; inoltre, mentre LexIt è costruito in modo completamente automatico, SIMPLE è stato sviluppato manualmente («SIMPLE: A general framework for the development of multilingual lexicons»).

mento delle preferenze di selezione agli argomenti dei verbi sia come filler lessicali che come classi semantiche, identificazione del ruolo semantico, estrazione automatica delle classi verbali. Il prodotto risultante è una risorsa per lo studio di 3.933 verbi italiani, un database comprendente frame sintattici, filler lessicali e classi semantiche per l'analisi statistica dei verbi italiani. LexIt può anche essere considerato come un lessico di valenza, poiché fornisce, per ciascun verbo, i più significativi pattern sintattici in termini di forza di associazione. La differenza principale tra Lexit e i tradizionali lessici di valenza, quali REDES (Bosque) e il Wörterbuch der italienischen Verben (Blumenthal e Rovere),[3] consiste nell'essere basato esclusivamente su corpus e costruito in modo completamente automatico. Nella costruzione di Lexit, il corpus scelto è *La Repubblica* (Baroni et al.), sviluppato all'SSLMIT dell'Università di Bologna. Il corpus consta di una collezione di circa 600.000 articoli pubblicati tra il 1985 e il 2000 dal quotidiano *La Repubblica*, per un totale di 386 milioni di parole, divise per tipologia e dominio. La scelta è ricaduta su *La Repubblica* per una serie di motivi: innanzitutto è considerato il corpus di riferimento per la lingua italiana, poi copre un lungo arco di tempo e il linguaggio usato è quello giornalistico, il quale dovrebbe essere rappresentativo dell'italiano standard; inoltre le categorizzazioni di tipo e dominio si prestano ad analisi ulteriori su specifici sottocorpora. Nella fase di preparazione vengono effettuate delle analisi automatiche: il corpus viene prima lemmatizzato e sottoposto a un Pos Tagging tramite ILC-UniPi Tagger (Dell'Orletta), poi a un *parser* a dipendenze stocastico, DeSR (Bosco, Montemagni e Mazzei). Dal corpus risultante viene ricavato un profilo distribuzionale per ogni verbo trattato e ogni profilo ottenuto viene organizzato in un

[3] *Il Wörterbuch der italienischen Verben* è l'unico lessico di valenza esistente per i verbi italiani, è basato su un corpus di 50 milioni di parole estratti da "Il Sole 24 Ore" (1989-1990). Il lessico è strutturato in modo che ogni lemma sia articolato in una serie di sensi, ciascuno correlata di esempi.

profilo sintattico e uno semantico. La metodologia usata per costruire LexIt è basata sulle tradizionali misure di associazione applicate allo studio delle collocazioni per valutare la forza di correlazione tra:

- verbi e frame sintattici (estrazione di frame di sottocategorizzazione);
- argomenti verbali e parole che li compongono (identificazione del set lessicale per lo slot di un frame sintattico);
- argomenti verbali e classi semantiche assegnate loro dall'algoritmo.

Profili distribuzionali e semantici per i verbi italiani

Il *distributional profile* di un verbo consiste nell'insieme di informazioni estratte da un corpus per caratterizzare le proprietà distribuzionali del verbo. In letteratura sono stati proposti numerosi metodi per l'acquisizione automatica dei dati, per l'estrazione di frame di sottocategorizzazione, per l'identificazione delle preferenze di selezione verbali (Manning e Schütze; Light e Greiff). Per quanto riguarda l'estrazione di frame di sottocategorizzazione[4] (SCFs) la caratteristica distintiva di Lexit è il suo approccio automatico e non supervisionato al problema: non si fornisce al modello una lista precostituita di frame sintattici, ma vengono identificati automaticamente le costruzioni sintattiche più frequenti nel corpus . Dopo il processo di estrazione si definiscono gli SCFs del modello. La lista dei suoi frame di sottocategorizzazione (SCFs) di un verbo, ordinati

[4]Per sottocategorizzazione verbale, o valenza verbale si fa riferimento alla capacità dei verbi di scegliere i propri complementi. La struttura di sottocategorizzazione può anche essere definita frame di sottocategorizzazione (SCF), questa fornisce uno strumento per formulare generalizzazioni sul comportamento dei verbi.

secondo la loro rilevanza statistica, definisce il suo profilo. Ogni SCF corrisponde a uno specifico pattern di dipendenze sintattiche di quello specifico verbo ed è formato da un set di slot e identificato da un'etichetta sintetica, ad esempio:

soggetto + complemento introdotto dalla *a* + oggetto diretto = *subj#obj#comp-a*

complemento introdotto dalla preposizione a + oggetto diretto = *comp-a#obj*

Nel modello viene anche considerato il pronome riflessivo *si* e il caso in cui un verbo appaia nella forma senza dipendenze, ad esempio nella frase "Il vaso si è rotto" (*subj#si#0*). Il processo di definizione del profilo sintattico avviene selezionando un numero di SCFs tra le combinazioni più frequenti, per ogni verbo si confronta la joint frequency con ogni SCF, basato su pattern estratti automaticamente dal corpus una volta sottoposto al *parser*. Dalla frequenza combinata verbo-SCF si ottiene il punteggio di Local Mutual Information[5] (LMI) che restituisce una stima della rilevanza statistica di quel SCF per il verbo dato. Nella costruzione di Lexit si parte dal metodo di Resnik per la procedura del conteggio della frequenza e dell'applicazione di una misura di associazione per valutare la correlazione tra un argomento e la classe semantica che occorre insieme ad esso, ma non si utilizzano tutti i synset di WordNet, bensì un gruppo di topnode selezionati che permette di avere una visione generalizzata dei set lessicali. Per l'assegnazione di profili semantici a questi slot sono state implementate una serie di variazioni all'algoritmo di Schulte Im Walde, tra le quali la divisione uniforme tra i sensi assegnati ai nomi nella sezione italiana di MultiWordNet[6] delle frequenze di

[5]La Local Mutual Information è una variante della Pointwise Mutual Information, che risolve alcuni problemi relativi agli eventi poco frequenti che venivano sottostimati. È una misura comunemente usata nell'analisi delle collocazioni lessicali.

[6]MultiWordNet è un lessico computazionale basato su WordNet. La sezione italiana è allineata con quella inglese: i synset italiani sono creati come corrispondenti di quelli in lingua inglese e le relazioni semantiche sono importate dal database inglese (Pianta, Bentivogli e Girardi).

co-occorrenza di ciascun nome nel ruolo di filler di un dato verbo e il calcolo del valore di associazione effettuato tra ogni combinazione verbo-frame-slot e le top-classi di WordNet. L'approccio usato nella costruzione di LexIt condivide dunque con Schulte Im Walde la selezione di un sottoinsieme di topnode esclusivi usati per un'analisi automatica, mentre le misure di associazione della relazione tra classi semantiche e predicati si basano sul modello precedente di Resnik. I profili semantici assolvono una funzione descrittiva e predittiva: da una parte i set lessicali forniscono una panoramica dei nomi che occorrono nel corpus con un verbo in una determinata posizione sintattica, con una valutazione della loro rilevanza statistica. D'altra parte le preferenze di selezione permettono di generalizzare a partire dalle istanze a delle proprietà astratte degli argomenti dei verbi, consentendo di formulare delle predizioni sugli argomenti non considerati. Un problema centrale in questa fase della costruzione del lessico è stabilire il livello di granularità dell'informazione semantica associata ai filler da MultiWordNet: nella prima fase si selezionano tutte le parole associate ai 24 topnode selezionati, ma questa operazione può creare delle anomalie nei risultati, derivanti dal fatto che MultiWordNet contiene termini anche molto specifici: ad esempio la parola *libro* non rientra solo nella tipologia Artifact e Communication, ma designa anche una parte dello stomaco dei ruminanti, quindi Body Part. Pertanto *libro* ha come associazione principale il verbo *leggere*, mentre Body Part è identificata come seconda classe maggiormente associata nello slot di oggetto diretto. Inoltre dalle analisi qualitative viene evidenziato come l'algoritmo non rappresenta correttamente gli usi metaforici, come *leggere la mano* o *leggere le labbra*.

Descrizione dell'esperimento

Il programma realizzato per l'esperimento mira a estrarre le preferenze di selezione dei verbi italiani basandosi sulle informazioni estratte da LexIt e utilizzando la gerarchia di MultiWordNet. LexIt, come detto precedentemente, contiene i profili distribuzionali dei verbi italiani estratti dal corpus "La Repubblica" e *Wikipedia*. Il formato dei dati di output estratti da LexIt utilizzati in questo lavoro è così composto: per ogni verbo sono specificati i frame di sottocategorizzazione, gli slot del frame, i filler nominali e le frequenza per ogni filler. Il formato dei dati estratti da LexIt è dunque composto dalla combinazione verbo-frame-filler-frequenza, come mostrato nell'esempio seguente.

testa verbale + frame	***filler* nominale**	**frequenza**
`avanzare-v%obj%obj`	`comitiva-s`	1
`avanzare-v%obj%subj`	`invasione-s`	2
`avanzare-v%si#0%subj`	`candidatura-s`	5

Tabella 1: Output LexIt filler

Tali dati estratti da LexIt forniscono precise informazioni lessicali, ma è necessaria una risorsa per generalizzare l'informazione relativa alle preferenze di selezione e associarla alla descrizione dei verbi rispetto a determinati argomenti. WordNet è stato ampiamente utilizzato come fonte per un'informazione a livello abbastanza dettagliato sulle preferenze di selezione (Ribas; Clark e Weir). Pertanto come fonte per l'informazione sulle preferenze di selezione viene utilizzato MultiWordNet che contiene anche la lingua italiana. Come detto sopra, MultiWordNet è organizzato in una gerarchia di synset sinonimi. La figura 2 fornisce una rappresentazione semplificata della gerarchia di MultiWordNet per il nome *salame*.

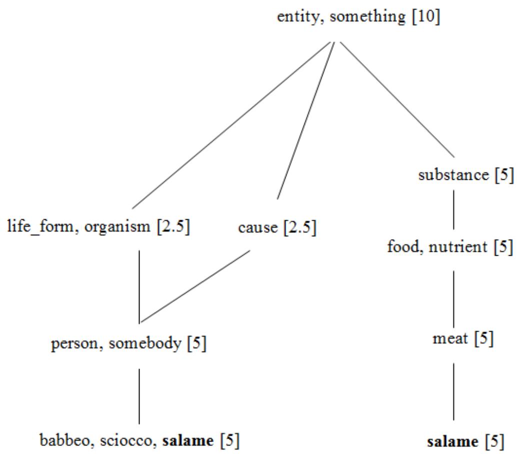


Figura 2: Gerarchia MWN

Come si vede dalla figura 2, il nome "salame" è associato a due sensi: *salame* come sinonimo di persona sciocca e *salame* come tipo di cibo. La gerarchia di MultiWordNet per ogni nome contenuto nei filler dei file estratti da LexIt, viene utilizzata per la costruzione delle preferenze di selezione per quella determinata combinazione verbo-frame-filler a quello specifico livello della gerarchia. L'approccio seguito nella costruzione del programma è simile a quello proposto da Schulte Im Walde («GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering»), e procede in questo modo. Per ogni nome in una combinazione verbo-frame-filler la frequenza è divisa per il numero di sensi del nome e propagata per tutta la gerarchia. Se un synset è collegato a iperonimi multipli, la frequenza viene divisa per il numero di iperonimi, se più synset puntano ad un unico iperonimo, la frequenza viene sommata. Ovviamente la somma

della frequenza dei topnode sarà uguale alla frequenza iniziale della combinazione verbo-frame-filler. Ad esempio se la frequenza del nome *salame* rispetto al verbo *mangiare* in posizione di oggetto diretto è uguale a 10, si assegna a entrambi i synset contenenti il nome un valore pari a 5. I valori di questi nodi si propagano per tutta la gerarchia, dividendosi in caso di nodi multipli e sommandosi in caso più nodi puntino a uno solo. Ripetere l'assegnamento della frequenza per tutti i nomi contenuti negli slot, fornisce una distribuzione della frequenza di tutte le combinazioni verbo-frame-filler estratte da LexIt su tutti i synset di MultWordNet. La novità nell'approccio proposto rispetto all'algoritmo di Schulte Im Walde consiste nel non esprimere le preferenze di selezione soltanto nei termini di una selezione di top-nodes mutualmente esclusivi, ma per tutti i nodi intermedi appartenenti alla gerarchia. Ricapitolando i passi seguiti nell'esecuzione del programma sono i seguenti:

1. per ogni nome in una combinazione verbo-frame-filler la frequenza del nome viene divisa per il numero dei suoi sensi;

2. questa operazione viene propagata a tutta la gerarchia dei synset, se un synset ha più iperonimi, la frequenza viene divisa per il numero di iperonimi, se più synset puntano a un solo iperonimo, la loro frequenza si somma;

3. l'operazione viene ripetuta per tutti i dati estratti da LexIt;

4. viene calcolata la forza di associazione usando la Local Mutual Information per ogni coppia verbo-frame rispetto a un dato synset.

Il formato di output dei dati è coerente con quello utilizzato da LexIt, ogni filler della sequenza verbo-frame-filler contenente un argomento nel file di partenza viene sostituito da tutti i synset che

incontra nel risalire la gerarchia, ad ognuno di questi viene associato il valore di frequenza corrispondente.

testa verbale + frame	synset	frequenza
avanzare-v%obj%obj	accomplishment	7.616
avanzare-v%obj%obj	accusation;accusal	23.375
avanzare-v%obj%obj	activity	1004.9
avanzare-v%si#0%subj	abstraction	70.993
avanzare-v%si#0%subj	act;human_action	218.25
avanzare-v%si#0%subj	affair;occasion	0.3333

Tabella 2: Output LexIt synset

Un elemento importante da considerare è che la lingua di partenza dei synset è l'italiano, in modo da potersi interfacciare con i dati estratti da LexIt, nella navigazione dell'albero si passa alla gerarchia dei synset inglesi, questa scelta è motivata dal voler usare la lingua inglese come metalinguaggio per rappresentare le classi, questo è reso possibile dal fatto che in MultiWordNet i concetti rappresentati sono organizzati in maniera parallela tra le varie lingue, come detto precedentemente.

Analisi dei risultati

Una volta navigata l'intera gerarchia di MultiWordNet i dati risultanti composti da verbo-frame-synset e rispettiva frequenza, vengono valutati utilizzando la Local Mutual Information per stabilire la forza di associazione tra quella coppia verbo-frame e quello specifico synset. Ottenuti i dati ordinati, si procede ad un'analisi qualitativa al fine di mettere in luce fenomeni che non sarebbe stato possibile osservare esprimendo le preferenze di selezione solo in termini dei top-node, come viene fatto attualmente in LexIt. Analizzare

le preferenze di selezione per ogni synset della gerarchia permette infatti di avere una rappresentazione più specifica del comportamento dei verbi nella selezione degli argomenti, basandosi sull'assunzione che le classi ai livelli intermedi della gerarchia possano caratterizzare meglio le specificità del comportamento dei verbi, impossibili da vedere considerando solo le classi in cima alla gerarchia (ovvero i top-nodes), troppo ampie e generiche. Un esempio che dimostra i risultati ottenuti dall'analisi riguarda il verbo *abbagliare* con frame di sottocategorizzazione *#obj* (oggetto diretto), il synset con un più alto maggiore di LMI risulta essere *radiation* nello slot. La tabella seguente e le successive rappresentano i dati nel formato del database di LexIt: verbo % frame di sottocatorizzazione % slot sintattico – synset associato alla fine dell'analisi e forza di associazione.

verbo-frame	**synset**	**LMI**
abbagliare-v%obj%subj	radiation	10.586

Tabella 3: Analisi risultati verbo: abbagliare

Se si guardasse soltanto ai top-nodes della gerarchia di MultiWordNet, il verbo andrebbe associato a *natural_phenomenon*, un concetto molto più generico e astratto che può esprimere un insieme di concetti anche molto diversi da *radiation*. È dunque un fenomeno rilevante rispetto all'analisi effettuata solo sui top-nodes e caratterizza meglio il comportamento del verbo per quel determinato frame. Per lo stesso frame si sono riscontrati altri casi degni di nota, alcuni esempi significativi possono essere: il verbo *abbandonare* il cui valore di LMI per il synset *housing;lodging* è molto più alto di quello per *artifact o physical_object* (1.341) e lascia intendere una specificazione maggiore del verbo sugli argomenti preferiti, riferiti a contesti specifici.

Negli esempi seguenti, verbo *avanzare* e *fingere*, si vede chiara-

verbo-frame	**synset**	**LMI**
abbandonare-v%obj%obj	housing;lodging	6.305
abbandonare-v%obj%obj	work	5.696
abbandonare-v%obj%obj	duty	4.091
abbandonare-v%obj%obj	energy	3.768

Tabella 4: Analisi risultati verbo: abbandonare

mente come i synset col più altro valore di LMI cui si associa il verbo risultano molto più caratterizzanti dei corrispettivi top-nodes. In questo caso l'unico top-node cui convergono *proposition*, *inactiveness*, *obedience* e *submissiveness* è *abstraction*, che avrebbe fatto perdere molta informazione sulle preferenze dei due verbi.

verbo-frame	**synset**	**LMI**
avanzare-v%obj%obj	proposition	10.421
fingere-v%obj%obj	inactiveness;inactivity	371
fingere-v%obj%obj	obedience	221
fingere-v%obj%obj	submissiveness	217

Tabella 5: Analisi risultati verbi: avanzare, fingere

Cambiando il frame si osservano altri risultati interessanti, ad esempio restringendo il campo di analisi al frame subj#0 si possono notare altri risultati interessanti, come ad esempio il verbo *avanzare*, il cui secondo synset con maggior valore di LMI, dopo *person*, *individual*, *someone* risulta essere *leader*.

verbo-frame	**synset**	**LMI**
avanzare-v%0%subj	leader	253,24

Tabella 6: Analisi risultati verbo: avanzare

Guardando il comportamento del verbo in LexIt, si nota che la classe semantica con cui ha maggior forza di associazione è *Person* (73.7787) e l'iperonimo top-node di *leader* in MWN è *life_form,oganism*, quindi il valore alto di LMI risultante da questa analisi restringe il campo e specifica meglio il concetto all'interno di una classe molto più ampia e varia.

Conclusioni e possibili sviluppi

L'analisi effettuata ha permesso di ottenere un livello di generalizzazione più specifico per descrivere il comportamento dei predicati nella selezione dei loro argomenti, effettuare l'analisi e calcolare i risultati per tutte le classi della gerarchia ha consentito di evidenziare delle specificità nelle preferenze di selezione messe in luce dalle classi intermedie di MultiWordNet.

L'analisi, come detto sopra, è effettuata navigando la gerarchia di MWN a partire dai synset italiani estratti da LexIt e passando ai livelli successivi con le classi della gerarchia inglese, perfettamente corrispondente alla gerarchia dei synset italiani. L'esperimento è stato implementato lavorando sui file di testo estratti da LexIt nel formato sopra desctitto (verbo-frame-frequenza) e i dati di output per esprimere le preferenze di selezione sono stati mantenuti nello stesso formato (verbo-frame-synset-frequenza), sostituendo al filler nominale il synset della classe corrispondente.

Nonostante questo lavoro possa considerarsi come un esperimento autonomo, indipendente da LexIt, qui utilizzato soltanto come risorsa lessicale-semantica per la distribuzione del comportamento dei verbi italiani, in un'ottica più ampia l'obiettivo e l'evoluzione naturale del programma sono quelli di integrarsi con le funzionalità di LexIt. Un'interrogazione diretta del database di LexIt e la possibilità di arricchirlo con tutte le informazioni fornite dall'analisi delle

preferenze di selezione su tutti i livelli della rete semantica, sono da considerarsi i possibili futuri sviluppi.

Riferimenti bibliografici

Baroni, Marco, et al. «Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian». *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*. Lisbon: ELDA, 2004. 1771–1774. (Cit. a p. 8).

Blumenthal, Peter e Giovanni Rovere. *Wörterbuch der italienischen Verben*. Stuttgart: Ernest Klettverlag, 1998. (Cit. a p. 8).

Bosco, Cristina, Simonetta Montemagni e Alessandro Mazzei. «Evalita'09 Parsing Task: comparing dependency parsers and treebanks». *Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence Reggio Emilia*. 2009. (Cit. a p. 8).

Bosque, Ignacio. *Redes: diccionario combinatorio del espanol contemporáneo*. Madrid: SM Ediciones, 2004. (Cit. a p. 8).

Chomsky, Noam. *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT press, 1965. (Cit. a p. 3).

Clark, Stephen e David Weir. «Class-Based Probability Estimation using a Semantic Hierarchy». *Computational Linguistics* 28. (2002). (Cit. a p. 12).

Cove, Thomas M. e Joy A. Thomas. *Elements of information theory*. New York: Wiley, 1991.

Dell'Orletta, Felice. «Maximum Entropy for Italian PoS Tagging». *Intelligenza Artificiale* 4.2. (2007): 10–11. (Cit. a p. 8).

Francis, Nelson e Henry Kucera. *Frequency analysis of English usage*. Boston: Houghton Mifflin, 1982. (Cit. a p. 4).

Gamallo, Pablo, Alexandre Agustini e Gabriel P. Lopes. «Clustering syntactic positions with similar semantic requirements». *Computational Linguistics* 31.1. (2005): 107–146.

Grefenstette, Gregory. *Explorations in automatic thesaurus discovery*. Boston: Springer, 1994.

Hamp, Birgit e Helmut Feldweg. «GermaNet: a Lexical Semantic Net for German». *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. 1997. (Cit. a p. 5).

Katz, Jerrold J. e Jerry A. Fodor. «The structure of a semantic theory». *Language* 39.2. (1963): 170–210. (Cit. a p. 3).

Kullback, Solomon e Richard A. Leibler. «On information and sufficiency». *Annals of Mathematical Statistics* 22. (1951): 79–86.

Lenci, Alessandro. «Carving Verb Classes from Corpora». *Word Classes*. A cura di Raffaele Simone e Francesca Masini. Amsterdam-Philadelphia: John Benjamins, 2010. (Cit. a p. 7).

—. «SIMPLE: A general framework for the development of multilingual lexicons». *International Journal of Lexicography* 22.4. (2000): 489–495. (Cit. a p. 7).

Light, Marc e Warren Greiff. «Statistical Models for the Induction and Use of Selectional Preferences». *Cognitive Science* 26. (2002): 269–281. (Cit. a p. 9).

MacWhinney, Brian e Catherine Snow. «The child language data exchange system». *Journal of Child Language* 12. (1985). (Cit. a p. 4).

Magnini, Bernardo e Carlo Strapparava. «Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet». *Atti del XXVIII Congresso della Società di Linguistica Italiana*. 1994. 415–418. (Cit. a p. 1).

Manning, Christopher e Hinrich Schütze. *Foundations of Statistical Language Processing*. Cambridge, MA: The MIT press, 1999. (Cit. a p. 9).

Pianta, Emanuele, Luisa Bentivogli e Christian Girardi. «MultiWordNet: developing an aligned multilingual database». *Proceedings of the First International Conference on Global WordNet*. 2002. (Cit. a p. 10).

Resnik, Philip. «Selection and information: a class-based approach to lexical relationships». phd. University of Pennsylvania, 1993.

—. «Selectional constraints: an information-theoretic model and its computational realization». *Cognition* 61. (1996): 127–159.

—. «Selectional preference and sense disambiguation». *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*. 1997. 52–57.

—. «Semantic classes and syntactic ambiguity». *Proceedings of the workshop on Human Language Technology*. 1993. 278–283. (Cit. a p. 1).

Ribas, Francesc. «On Learning More Appropriate Selectional Restrictions». *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*. 1995. (Cit. a p. 12).

Schulte Im Walde, Sabine. «Experiments on the Automatic Induction of German Semantic Verb Classes». *Computational Linguistics* 32.2. (2006): 159–194. (Cit. a p. 2).

—. «GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering». *Proceedings of the GermaNet Workshop*. 2003. (Cit. a p. 13).

RAFFAELE GUARASCI, Università degli Studi di Pisa.
r.guarasci@studenti.unipi.it

Guarasci, R."Le preferenze di selezione verbali: un approccio computazionale". *JLIS.it* Vol. 3, n. 1 (Giugno/June 2012): 4786-1–4786-21. DOI: 10.4403/jlis.it-4786. Web.

ABSTRACT: Il lavoro mira a fornire una rappresentazione delle preferenze di selezione verbali per la lingua italiana. L'esperimento si ricollega alle metodologie basate su corpora e si articola in due fasi: l'estrazione degli argomenti dai corpora scelti e la generalizzazione delle preferenze di selezione utilizzando un'ontologia lessicale. Le risorse utilizzate sono: LexIt, un lessico di valenza per i verbi italiani, come risorsa lessicale, e MultiWordNet, come ontologia. L'obiettivo è fornire un livello di rappresentazione dettagliato del comportamento verbale navigando l'intera rete semantica e facendo emergere comportamenti più specifici nelle preferenze di selezione degli argomenti verbali.
KEYWORDS: Linguistica computazionale; LexIt; MultiWordNet; Selezione verbale; Verbi

---

Submission: 2012-02-12
Accettazione: 2012-03-02
Pubblicazione: 2012-06-01